ОЦЕНКА ПОГРЕШНОСТЕЙ РАНДОМИЗИРОВАННОГО АДГОРИТМА ВОССТАНОВЛЕНИЯ ПРОПУСКОВ ВО ВРЕМЕННЫХ РЯДАХ ДИСТАНЦИОННЫХ ИЗМЕРЕНИЙ ПЛОЩАДЕЙ ОЗЕР АРКТИКИ

Гл. специалист Сокол Е.С. [1] , д.т.н., проф. Попков Ю.С. [2], д.т.н., проф. Мельников А.В. [1], к.т.н. Полищук В.Ю. [3], д. ф.-м.н., проф. Полищук Ю.М. [1]

[1]АУ «Югорский НИИ Информационных Технологий», Ханты-Мансийск
[2]Федеральный исследовательский центр "Информатика и управление" РАН, Москва,
[3]ФГБУН «Институт мониторинга климатических и экологических систем СО РАН», Томск

Для восстановления пропущенных значений во временных рядах площадей озер, связанных с недостаточным числом безоблачных спутниковых снимков из-за высокой степени облачности на арктических территориях, предложен рандомизированный алгоритм восстановления пропусков в экспериментальных данных о площадях термокарстовых озер с использованием временных рядов среднегодовой температуры и годовой суммы осадков, основанный на методах энтропийно-рандомизированного машинного обучения. В качестве экспериментальных данных о площадях термокарстовых озер и климатических параметрах (температура и сумма осадков) использованы результаты исследований, проведенных на девяти тестовых участках в Арктической зоне Западной Сибири в период с 1973 по 2007 г. Проведен анализ погрешностей восстановления пропусков, который показал, что рандомизированный алгоритм позволяет восстанавливать пропущенные значения площадей озер с использованием временных рядов температуры и осадков с практически приемлемой точностью.

Работа проводилась в рамках проекта по госзаданию Департамента информационных технологий и цифрового развития Ханты-Мансийского автономного округа и частично при поддержке грантов по проектам РФФИ № 18-47-700001 и № 19-07-00282.

Введение

Особую важность рандомизированный подход представляет для решения задач прогнозирования динамики накопления парниковых газов в термокарстовых озерах арктической зоны в связи с их влиянием на глобальные климатические изменения, что может явиться основой разработки и функционирования систем адаптации к меняющимся условиям среды обитания на различных управленческих уровнях. В условиях грядущего глобального потепления в ближайшие десятилетия будут ускоряться процессы таяния мерзлых пород, приводя к дополнительному высвобождению углекислого газа и метана - парниковых газов, способных внести дополнительный ощутимый вклад в потепление климата. Значительный вклад в парниковый эффект вносят термокарстовые озера в зоне мерзлоты. В связи с необходимостью оценки этого вклада особую важность приобретает решение задач прогнозирования динамики накопления парниковых газов в термокарстовых озерах в ближайшие десятилетия, решение которых требует использования данных о временных рядах площадей озер и климатических параметров (температуры, осадков).

Ввиду большой степени заболоченности арктических территорий данные о площадях озер могут быть получены только с использованием спутниковых снимков. Большое количество пасмурных дней на северных территориях приводит к малому числу безоблачных снимков лишь в отдельные годы. В результате этого полученные временные ряды площадей озер обладают значительным числом пропущенных значений. Вопросы восстановления пропусков во временных рядах, применительно к данным о площадях озер, в настоящее время разработаны недостаточно. Наиболее перспективным подходом к восстановлению пропусков в наших условиях рассматривается использование энтропийно-рандомизированных методов, показавших, согласно [1-3], высокую эффективность в решении прогнозных задач глобальной экономики, демографии и др. Однако погрешности рандомизированного алгоритма восстановления пропусков во временных рядах площадей озер в настоящее время не исследованы, что и явилось целью настоящей работы.

Данные и методы

Информационную основу проведения исследований погрешности рассматриваемого алгоритма восстановления пропусков составляют экспериментальные данные об изменениях климатических параметров и размеров озер на территории Западно-Сибирской Арктики. В качестве экспериментальных данных о площадях термокарстовых озер и климатических параметрах (температура и сумма осадков) использованы результаты исследований, проведенных на девяти тестовых участках исследуемой территории в период с 1973 по 2007 г. Схема размещения тестовых участков (ТУ) для проведения исследований, направленных на получение данных о

временных рядах площадей термокарстовых озер, среднегодовой температуры и годовой суммы осадков приведена в [4].

Для получения данных о площадях озер использованы космические снимки среднего разрешения (30 м) Landsat, полученные в теплый период года (как правило, в июле-августе), когда отсутствует ледовый покров, мешающий автоматическому дешифрированию озер на космических снимках с использованием средств географической системы ArcGIS [5]. Данные о среднегодовой температуре и годовой сумме осадков для каждого тестового участка получены методом реанализа данных [6]. Для иллюстрации в [4] представлен в табличном виде массив данных о средней площади термокарстовых озер $\tilde{S}_{изм}$, среднегодовой температуре $\tilde{T}$, годовой сумме осадков $\tilde{R}$ на одном из тестовых участков ТУ-1. Подмассив данных, относящийся к периоду 2001-2007 гг., приведен для иллюстрации ниже в таблице 1.

Таблица 1.

Данные о площади, температуре и осадках на ТУ-1

Годы / Данные	1	2001	2002	2003	2004	2005	2006	2007
$\tilde{S}изм, га$	2	62,36	-	62,31	-	-	-	66,96
$\tilde{T}$, °С	3	-2	-2,06	-0,48	-1,8	0,44	-2,78	0,14
$\tilde{R}$, мм	4	430,2	520,15	429,2	338,4	343,9	222	325,4
$\tilde{S}_{восст}$,	5	62,36	65,21	62,31	62,9	62,07	61,86	66,96

Как видно из таблицы 1 (строка 2), данные о площадях озер имеют большое число пропущенных значений, что является характерным для всех исследованных тестовых участков. В строке 5 таблицы 1 представлены восстановленные (модельные) данные о средней площади озер. Заметим, что в таблице 1 в те годы, в которых имелись реальные измеренные значения площадей озер (строка 2), в строке 5 вместо восстановленных данных использованы измеренные значения. Восстановление пропущенных значений проводилось как расчет модельных значений площадей озер с использованием данных о среднегодовой температуре и годовой сумме осадков в соответствии с методологией моделирования динамики площадей озер в рамках энтропийно-рандомизированного подхода, предложенного в [7]. Поэтому алгоритм восстановления пропущенных данных о площади озер, основанный на этом подходе, будем называть рандомизированным.

Алгоритм восстановления пропусков достаточно подробно рассмотрен в [4]. Прежде чем им пользоваться, необходимо выполнить преобразование измеренных данных по площади $\tilde{S}$, температуре $\tilde{T}$, осадкам $\tilde{R}$ к стандартному (нормализованному) виду по следующим формулам:

$$S = \frac{\tilde{S} - \tilde{S}_{min}}{\tilde{S}_{max} - \tilde{S}_{min}},$$
$$T = \frac{\tilde{T} - \tilde{T}_{min}}{\tilde{T}_{max} - \tilde{T}_{min}}, \quad (1)$$
$$R = \frac{\tilde{R} - \tilde{R}_{min}}{\tilde{R}_{max} - \tilde{R}_{min}}.$$

где S, T, R – нормализованные значения площади, температуры и осадков, отображенные на промежутке [0,1]; нижний индекс у всех показателей означает минимальное и максимальное значение выборки.

В нашем случае временные ряды измеренных данных о площади озер обладают большим числом пропущенных значений. Для восстановления пропущенных данных воспользуемся принципом энтропийной рандомизации по площади, используя имеющиеся данные по температуре и осадкам. Известно, что на площадь озер влияет температура и осадки, и в первом приближении это влияние можно описать линейной зависимостью с шумом в виде:

$$S[n] = \alpha T[n] + \beta R[n] + \xi[n]. \quad (2)$$

Коэффициенты α, β - случайные, интервальные:

$$\alpha \in \mathcal{A} = [\alpha^-, \alpha^+], \quad \beta \in \mathcal{B} = [\beta^-, \beta^+]. \qquad (3)$$

Вероятностные свойства параметров модели (коэффициентов α, β) характеризуются функцией плотности распределения вероятности (ПРВ). Обозначим плотность распределения вероятностей (ПРВ) параметров $P(\alpha), F(\beta)$. Шум также стандартизованный и интервальный:

$$\xi[\mathrm{n}] \in \Xi_j = [\xi^-, \xi^+]. \qquad (4)$$

Обозначим ПРВ шума $Q_n(\xi[n])$. Далее, применяя алгоритм рандомизированного машинного обучения, получим соотношение для энтропии:

$$\mathcal{H} = -\int_{\mathcal{A}} P(\alpha) \ln P(\alpha)\, d\alpha - \int_{\mathcal{B}} F(\beta) \ln F(\beta)\, d\beta - \qquad (5)$$
$$-\sum_{m=1}^{k} \int_{\Xi_m} Q_m(\xi[m])\, ln\, Q_m(\xi[m]) d\xi[m] \Rightarrow max$$

при условии нормировки:

$$\int_{\mathcal{A}} P(\alpha)\, d\alpha = 1,$$
$$\int_{\mathcal{B}} F(\beta)\, d\beta = 1,$$
$$\int_{\Xi_m} Q_m(\xi[m]) d\xi[m] = 1, \quad m = \overline{1,k}$$

и эмпирических балансов:

$$\int_{\mathcal{A}} P(\alpha)\alpha T[m]\, d\alpha + \int_{\mathcal{B}} F(\beta)\beta R[m]\, d\beta +$$
$$+\int_{\Xi_m} Q_m(\xi[m])\xi[m] d\xi[m] = S[m], \quad m = \overline{1,k}.$$

Согласно [4], решение задачи (5) имеет вид:

$$P^*(\alpha, \theta) = \frac{exp(-\alpha\, l_r(\theta))}{\mathcal{P}(\theta)},$$
$$F^*(\beta, \theta) = \frac{exp(-\beta\, h_r(\theta))}{\mathcal{F}(\theta)}, \qquad (6)$$
$$Q_m^*(\xi[m], \theta) = \frac{exp(-\theta_m, \xi[m])}{Q_j(\theta_m)}, \qquad m = \overline{1,k}$$

где $\theta = \{\theta_1, \ldots, \theta_k\}$ - множители Лагранжа, определяемые в нашей работе по методике [4].

Используя модель для расчета всех пропущенных данных и сэмплируя ПРВ, построим ансамбль траекторий S[n]. Вычислим среднюю траекторию и по ней заполним недостающие данные. После этого преобразуем данные из нормализованных значений к натуральным величинам площади озер (га). Для этого произведем действие, обратное к (1), а именно:

$$\tilde{S}_{\text{восст}} = S * (\tilde{S}_{max} - \tilde{S}_{min}) + \tilde{S}_{min}.$$

Результаты

Сравнительный анализ линейных трендов временных рядов. В соответствии с алгоритмом были восстановлены данные о площадях озер на всех тестовых участках. Для иллюстрации на рисунке 1 представлены результаты восстановления пропусков в виде графиков временных ходов средних (по всем ТУ) нормализованных значений восстановленных и измеренных данных о площади озер и линий трендов этих временных рядов. Как видно из рисунка, линии трендов лишь незначительно расходятся, что позволяет судить о приемлемом качестве предложенного рандомизированного алгоритма восстановления пропусков с использованием климатических параметров. Для количественной оценки согласованности линейных трендов восстановленных и измеренных данных в таблице 2 приведены значения коэффициентов трендов временных рядов, полученных на разных ТУ.

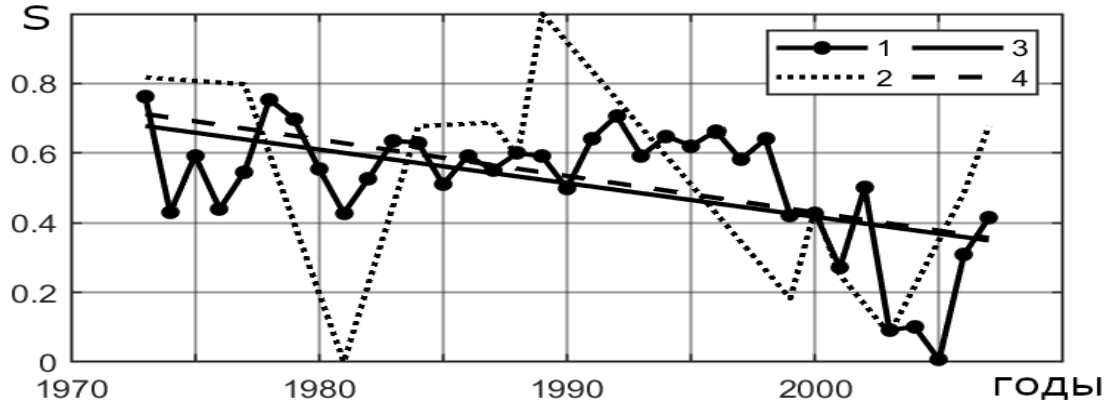

Рисунок 1. Временные ряды средних значений нормализованных восстановленных (1) и измеренных (2) данных о площади озер по всем ТУ, линии тренда временных рядов средних значений восстановленных (3) и измеренных (4) данных о площади озер

Таблица 2.

Коэффициенты линейных трендов временных рядов восстановленных и измеренных данных

Номер ТУ	ТУ-1	ТУ-2	ТУ-3	ТУ-4	ТУ-5	ТУ-6	ТУ-7	ТУ-8	ТУ-9
Восстановленные данные	-0,0134	-0,0102	-0,0082	-0,0100	-0,0045	0,0022	-0,0429	0,0026	-0,0021
Измеренные данные	-0,0192	-0,0096	-0,0021	-0,0440	-0,0085	-0,0166	-0,0231	0,0013	-0,0114

Проведено сравнение коэффициентов трендов рядов измеренных и восстановленных данных на разных ТУ. Были получены следующие значения для временных рядов восстановленных данных: максимальный коэффициент тренда 0,0026, минимальный -0,0429, медианное значение -0.0082 и аналогично - для трендов измеренных данных: максимальный коэффициент тренда 0,0013, минимальный -0,0440, медианное значение -0,0114. Приведенные данные показывают близость полученных оценок.

Анализ точностных показателей алгоритма. Оценка погрешности восстановления пропусков во временных рядах проводилась как оценка среднего отклонения восстановленных значений от измеренных данных. Были использованы измеренные и восстановленные значения на всех тестовых участках в те годы, для которых были получены измеренные данные о площадях озер. Для каждого тестового участка рассчитаны средние отклонения восстановленных данных от измеренных значений по следующей формуле:

$$\Delta_s = \frac{1}{m}\sum_{i=1}^{m}\frac{|\tilde{S}_{восст_i} - \tilde{S}_{изм_i}|}{\tilde{S}_{изм_i}},$$

где m – количество восстановленных значений на конкретном ТУ.

Полученные данные приведены в таблице 3.

Таблица 3.

Погрешности восстановления данных

ТУ	1	2	3	4	5	6	7	8	9	Среднее значение
Δ_s	0,04	0,15	0,08	0,04	0,04	0,06	0,04	0,07	0,04	0,05

Кроме этого были рассчитаны величины среднего и среднеквадратического отклонения восстановленных значений площади озер от измеренных величин для всего массива данных, т.е. на всех тестовых участках. Расчет среднего отклонения проводился по формуле:

$$\Delta = \frac{1}{k}\sum_{i=1}^{k}\frac{|\tilde{S}_{восст_i} - \tilde{S}_{изм_i}|}{\tilde{S}_{изм_i}},$$

где k – количество всех измеренных значений временного ряда по всем тестовым участкам.

В нашем случае было получено: $\Delta = 0{,}06$ и среднеквадратическое отклонение, рассчитанное по стандартной формуле, равно 0,09. Следовательно, погрешность восстановления пропусков по

разработанному алгоритму в среднеквадратическом не превышает 9%, что можно рассматривать как практически приемлемый результат.

Литература

1. Попков Ю.С., Попков А.Ю., Дубнов Ю.А. Рандомизированное машинное обучение при ограниченных объемах данных. – М.: УРСС, 2019. – 310 с.
2. Иоффе А.Д., Тихомиров В. Теория экстремальных задач. – Москва: Наука, 1974. – 435 с.
3. Popkov Y.S., Popkov A.Y. New Method of Entropy-Robust Estimation for Ramdomized Models under Limited Data // Enthropy, 2014. Vol. 16. P. 675-698.
4. Попков Ю.С., Мельников А.В., Полищук Ю.М., Сокол Е.С., Полищук В.Ю. Новый подход к восстановлению пропущенных данных о площади термокарстовых озер Арктики // Труды 8-й Междун. науч. конф. «Информационные технологии и системы» (Ханты-Мансийск, 17- 21 марта 2020 г.) / отв. ред. Ю.С. Попков и А.В. Мельников. Науч. электрон. изд-е [Электронный ресурс]. - Ханты-Мансийск: 2020. С. 32-39.
5. Полищук В.Ю., Полищук Ю.М. Геоимитационное моделирование полей термокарстовых озер в зонах мерзлоты. – Ханты-Мансийск: УИП ЮГУ, 2013. – 129 с.
6. Polishchuk Y.M., Muratov I.N., Polishchuk V.Y. Remote research of spatiotemporal dynamics of thermokarst lakes fields in Siberian permafrost. Chapter 8 In: The Arctic: Current Issues and Challenges (Eds. O.S. Pokrovsky, S.N. Kirpotin and A.I. Malov). New York: Nova Science Publishers, 2020. P. 208-237.
7. Попков Ю.С., Волкович З., Мельников А.В., Полищук Ю.М. Методологические вопросы использования рандомизированного машинного обучения для прогнозирования динамики термокарстовых озер Арктики // Вестник Южно-Уральского государственного университета. Сер. «Компьютерные технологии, управление, радиоэлектроника». 2019. Том 19. № 4. С. 5-12.

EVALUATION OF ERRORS OF A RANDOMIZED ALGORITHM OF RESTORING MISSING DATA IN TIME SERIES OF REMOTELY MEASURED LAKES AREAS IN THE ARCTIC

Sokol E.S. [1] , Popkov Y.S. [2], Melnikov A.V. [1], Polishchuk V.Y. [3], Polishchuk Y.M. [1]

[1]Ugra Research Institute of Information Technologies, Khanty-Mansiysk, Russia,
[2]Federal Research Center “Informatics and Control”, RAS, Moscow, Russia,
[3]Institute for Monitoring of Climatic and Ecological Systems, SB RAS, Tomsk, Russia

To restore missing values in the time series of lake areas associated with an insufficient number of cloudless satellite images due to the high degree of cloudiness in the Arctic territories, a randomized algorithm for restoring missing values in the experimental data on the areas of thermokarst lakes using time series of average annual temperature and annual precipitation was proposed. The algorithm is based on methods of entropy-randomized machine learning. As experimental data on the areas of thermokarst lakes and climatic parameters (temperature and amount of precipitation), we used the results of studies conducted at nine test sites in the Arctic zone of Western Siberia during the period from 1973 to 2007. An analysis was made of the errors in the restoration of missing data, which showed that a randomized algorithm allows to restore missing values of lake areas using time series of temperature and precipitation with practically acceptable accuracy.